Decoders Matter for Semantic Segmentation: Data-Dependent Decoding Enables Flexible Feature Aggregation

Zhi Tian[1] Tong He[1] Chunhua Shen[1*] Youliang Yan[2]
[1]The University of Adelaide, Australia [2]Noah's Ark Lab, Huawei Technologies

CVPR2019

目的：

提出数据依赖上采样(DUpsampling)取代双线性上采样，从低分辨率输出中恢复像素级预测，同时大量降低计算的复杂性。

方法：

大部分解码器的特征融合都是先将深层的低分辨率输出上采样后再与浅层特征图合并，然后再使用双线性上采样恢复分辨率。而这种方式的运算量较大且双线性上采样没有很好利用数据的关系。作者认为像素的标签并不是独立的，包含着结构信息也就是存在关联，因此可以压缩 label 而且能够依赖这种结构信息重构 label 而不会有太多的损失。

$$\boldsymbol{x} = \mathbf{P}\boldsymbol{v}; \quad \tilde{\boldsymbol{v}} = \mathbf{W}\boldsymbol{x},$$

用矩阵 P 来压缩向量，W 解码，使得向量 v 在压缩与重建之后尽可能相似。

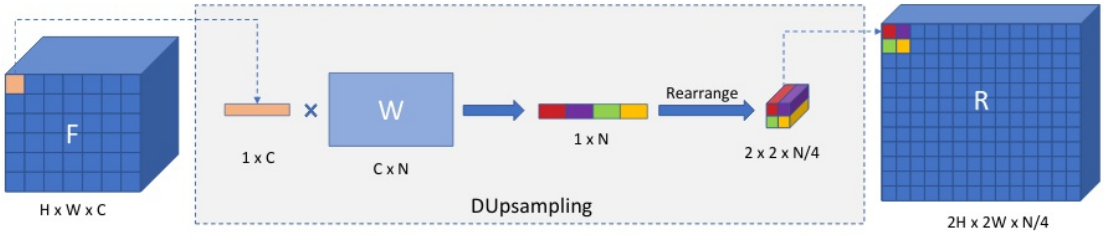
$$\begin{aligned}\mathbf{P}^*, \mathbf{W}^* &= \arg\min_{\mathbf{P},\mathbf{W}} \sum_{\boldsymbol{v}} ||\boldsymbol{v} - \tilde{\boldsymbol{v}}||^2 \\ &= \arg\min_{\mathbf{P},\mathbf{W}} \sum_{\boldsymbol{v}} ||\boldsymbol{v} - \mathbf{W}\mathbf{P}\boldsymbol{v}||^2.\end{aligned}$$

P,W 的求解即上述的优化目标，可用梯度下降来优化。这可以看成是一个线性的自编码器。

而作者直接用一个矩阵 W 将低分辨率输出变换上采样然后与标签计算损失：

$$\mathrm{L}(\mathbf{F}, \mathbf{Y}) = \mathrm{Loss}(\mathrm{softmax}(\mathrm{DUpsample}(\mathbf{F})), \mathbf{Y}).$$

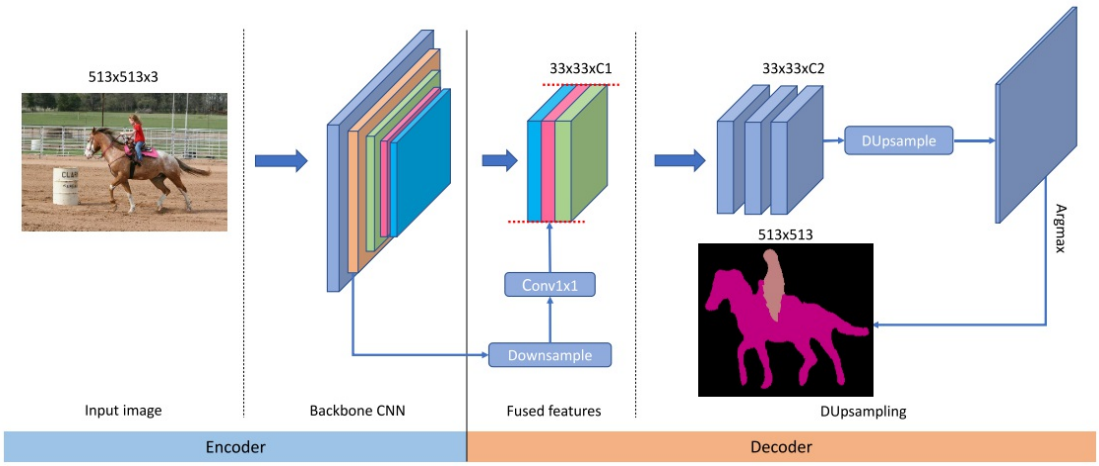
如下图所示：



DUpsampling 与 softmax 的这种组合难以产生尖锐的激活，也就是概率分布比较平滑，这样导致训练时的损失计算会卡住。产生这种现象的原因作者认为可能是因为 W 是根据 one-hot 的 label 计算得到的，因此为了解决这个问题引入了 Adaptive temperature Softmax 使得网络收敛更快：

$$\mathrm{softmax}(z_i) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}.$$

网络的结构参考 DeeplabV3+，在解码器部分做了改动：



总结：

对比 CARAFE 根据邻域重组得到新的像素，本文的方法是将新像素的信息与低分辨率特征图中的某一个像素联系起来，更加快速但是利用的信息也更少了，而且全局使用一个 W,对结构信息更加复杂的标签可能不太好，但是比双线性会好，毕竟引入了可学习参数。